


HKUST technologies and applications are enabling the massive amounts of information now being generated to drive forward business and social change

BIG DATA, DISRUPTIVE INNOVATION



In 1998, Google received around 10,000 search queries per day.* Now the search engine handles over 3.5 billion search queries per day or an average of 40,000 every second. Where decision-makers used to rely largely on experience, in today's mega-dataset digital era, huge amounts and varieties of information can be quickly accessed, integrated and analyzed to reveal fresh knowledge and quantitative patterns that widen established perspectives and norms. Traffic jam predictions for bad weather through combining meteorological and transport data, or analysis of Twitter messages discussing illness together with flight information to prevent the spread of diseases are examples.

Through exciting advances in this brave new world of colossal and

integrated data, HKUST computer scientists and engineers are playing a major role in bridging the traditional gap between academia and industry practice. The focus of the cutting-edge research at the University and through the HKUST Big Data Institute is to tackle problems with real-world relevance and to collaborate with industry front-runners to turn theory into applications leading to innovative services and new ways of understanding ourselves.

To further develop cognitive computing, researchers are fostering novel insights into artificial intelligence, machine learning, and computer visualization that are extending big data analytics into areas such as the arts and business writing (stylometry and machine reading), retail/consumer

recommendation systems (transfer learning), education (learner behavior), and smart city infrastructure (for example, route planning, visitor traffic).

To assist such breakthroughs, HKUST computer engineers are making significant contributions to big data infrastructure support through leading-edge improvements in speed and efficiency. These advances include developments in database query processing and interactive data exploration; and faster communication between machines in the physical data centers that keep search engines and cloud computing in operation.

In doing so, the University's researchers are helping both machines and humans become smarter.

* Statistic Brain Research Institute

BUILDING INTELLIGENCE

From Numbers to Knowledge

For a person, the ability to recognize and apply knowledge and skills learned in previous tasks to new endeavors is a natural occurrence. After understanding how one card game works, it is easier to pick up another. For a computer, such learning is incredibly hard. This is the specialist domain of Prof Qiang Yang, an expert in data mining, artificial intelligence, machine learning, transfer learning and deep reinforcement learning.

Prof Yang has spent 20 years fathoming algorithms that seek to endow computers with similar capabilities to humans in retaining and reusing previously learned knowledge in order to “think” and “decide” how to extract information and patterns from the rivers of data flooding our digital age. Prof Yang and his team have improved the accuracy of computers’ performance through devising versatile frameworks for such “transfer learning”. He has developed Instance-based Transfer Learning, which uses individual instances

to bridge different domains, and Heterogeneous Transfer Learning, where the computer learns in one knowledge domain (for example, text) then transfers what it has learned to a separate or more difficult domain (for example, images).

Prof Yang has made these frameworks open source, enabling other researchers and the field overall to develop at a faster pace. He was also the first to propose the use of transfer learning in collaborative filtering and recommender systems. Applications have ranged from early online advertising directed at users to improvements in recommendation systems, including a state-of-the-art recommendation system for ICT global giant Huawei’s App Store.

Recent research at the WeChat-HKUST Joint Lab on Artificial Intelligence Technology (WHAT LAB), set up with Mainland China internet giant Tencent in 2015, has inspired a novel application to improve machine reading capabilities. Books, news articles, and blogs are used as input to train a machine learning model that can produce an

“ We are inventors, always thinking of how to use data in a new way ”



PROF QIANG YANG

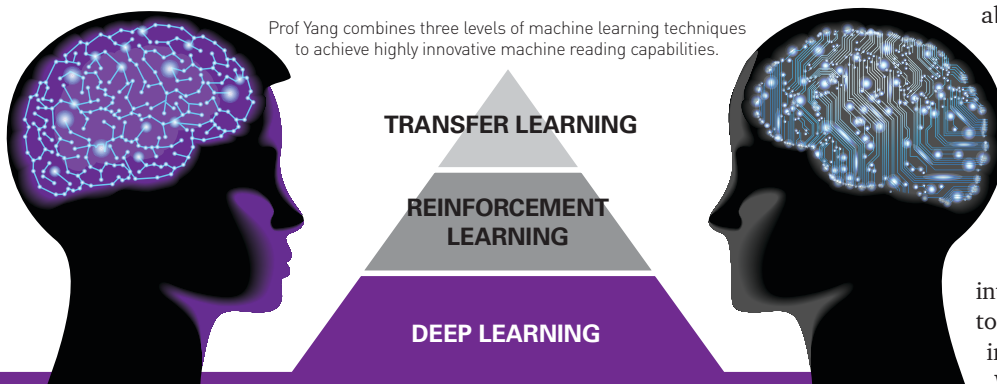
New Bright Professor of Engineering, Head, Department of Computer Science and Engineering, Director, HKUST Big Data Institute, Inaugural Editor-in-Chief of IEEE Transactions on Big Data

abstract of such readability that it doesn’t appear to have been written by a machine. The objective is to assist people with information overload on social networks or boost company productivity by enabling a computer to develop an abstract of a long report or integrate data and highlight the main points the reader needs to know. By reading books by the same author, the demonstration model designed by Prof Yang and his students has even written a high-quality novel of its own in the writer’s style, taking just a few seconds to do so.

In improving such machine reading abilities, Prof Yang’s team has become the first to integrate a reinforcement learning algorithm that leverages users’ feedback related to positive comments on prior abstracts with transfer learning and deep learning (recurrent neural networks) to help the computer make a more intelligent decision on what abstract to generate. The innovation has improved the quality substantially. With information to hand quicker, it can also speed up report-writing as well as learning.

“The work at HKUST seeks to increase the knowledge we can get from data by making the process of moving from data source to understanding faster and more efficient, accurate and useful to people,” Prof Yang said.

Prof Yang combines three levels of machine learning techniques to achieve highly innovative machine reading capabilities.



Humans have the ability to apply knowledge and skills learned in previous tasks (e.g. cycling) to new tasks (e.g. motorcycling).

Prof Yang’s transfer learning algorithms help computers to acquire the ability to retain and transfer knowledge from a source domain to a target domain.



“ All these charts and visuals are like a movie. The actors are the same, but when you combine them together differently, you can tell a new story ”



PROF HUAMIN QU
Professor of Computer Science and Engineering

Seeing the Larger Picture

The power of the visual to impart information plays a hugely significant part in our lives, shaping our understanding of the world through “seeing with our own eyes” and through a variety of media, ranging from art over the ages to today’s selfies and YouTube videos. Prof Huamin Qu and his team are leveraging such visual impact to mine the digital world of big data, by combining computational power to detect patterns and extract information from vast quantities of data with cutting-edge graphics and virtual reality techniques. In this way, they are uncovering previously unknown relationships, including those related to our own behavior. “We call it amplifying cognition,” Prof Qu said.

One recent outcome of such data visualization is VisMOOC, the first visual analysis system for discerning e-learning behavior. The intuitive HKUST web app offers fine-grained analysis of video “clickstream” data, drawn from learners

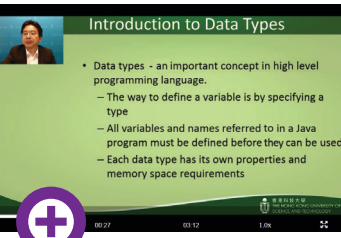
watching lectures for Massive Open Online Courses (MOOCs). VisMOOC pinpoints where learners play a section multiple times (indicating difficulty in comprehension), where they pause (to consider or take notes) and what they skip through (lack of interest or not challenging enough), among other details. Such clickstream data are matched with statistics from chat groups (forums), demographics, and grading for assignments and exams. Results are then provided in a novel visual form, labeled a “seek diagram”.

Following VisMOOC’s success, Prof Qu’s team and collaborators are developing an open source platform with advanced visualization interfaces for individual institutions to do detailed analysis on e-learning behavior and course design.

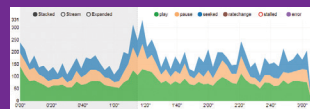
On a wider scale, Prof Qu is integrating cutting-edge visualization with large-scale telecommunications data to create applications contributing to smart city understanding, for example, route planning, crowd management for transportation, analysis of visitor traffic for shopping centers at different times of the day, and even tracking of disease outbreaks. In 2016, such work saw Prof Qu receive the Distinguished Collaborator Award from Huawei’s Noah’s Ark Lab, the company’s long-term, big-impact research lab. Working with WeChat, Mainland China’s dominant messenger app, Prof Qu has also solved the challenging problem of visualizing the propagation of information over a large social media network, involving multiple attributes/dimensions and dynamic evolution. Analyzing users’ behavior can assist in finding common communication patterns adopted by the public.

According to Prof Qu, a good visualization design must be effective in serving as a magnifying glass for what the data patterns show, aesthetic and intuitive. In addition, it should not be a pie chart or bar graph but a new visual form that carries interest for the viewer. Such integration of computational power in pattern recognition and mining and human expertise in visual pattern recognition, is a key area for further exploration, he noted. “Many real-world problems cannot be easily formulated as a computer algorithm so we need to keep humans in the loop.”

Visual Analysis for Massive Open Online Courses (MOOCs)



Seek Graph: orange lines are forward seeks, indicating that students skipped certain parts of the video; blue lines are backward seeks, meaning students went back to watch sections of the video. The thick blue lines indicate video sections watched multiple times, possibly to gain content clarity, thus alerting instructors to evaluate the course content and delivery.



Event Graph: showing six different types of clickstream data - play, pause, sought, rate change, stalled, and error - of the same course during the same time period but for students from different countries. By filtering demographic info on the dashboard, instructors can explore and compare online learning behaviors internationally.

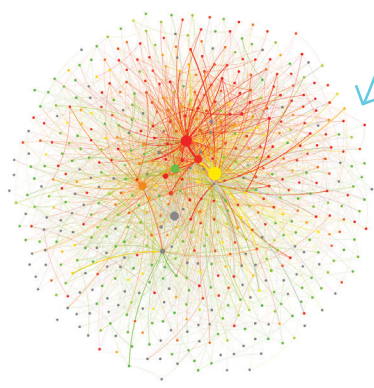
Dashboard

Course

- Course Info
- Popularity Info
- Age Info
- Demographic Info
- Animation
- Video
- Forum
- Sentiment plays
- Social Network
- Overview
- Correlation

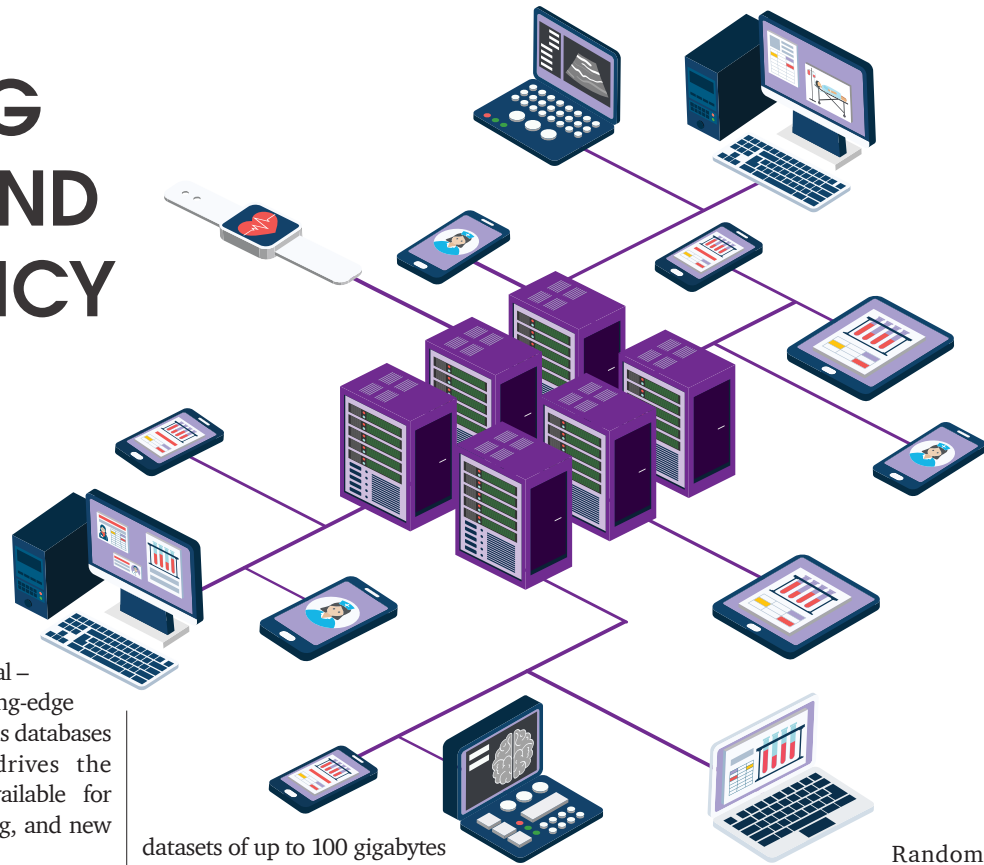
Student's grade: 15 (green) to 100 (red) | No grade (gray)

Student's activeness: Less (small dot) to More (large dot)



Forum Social Network: each dot represents a student. The size indicates the degree of activeness, and the color indicates the grade. The large gray dot shows that the student is active in the forum but does not achieve any score. The visualization gives valuable insight about students and their learning interactions, and provides MOOC instructors with useful information to improve course content.

BUILDING SPEED AND EFFICIENCY



Responding to Queries... Fast

Building intelligence – both human and computational – requires the support of leading-edge big data infrastructure, such as databases and data centers, that drives the fundamental capabilities available for data analysis, decision-making, and new ways of understanding.

Within the realm of digital databases, for example, the keywords are speed, accuracy, and accessibility. These are the criteria that count for data analysts and business users who need to perform analytical queries on a large amount of data with complex conditions and return aggregated results that can enable decisions to be made. For example, a sales director of a company handling millions of transactions per day might be interested to know the total revenue of all transactions for a specific product category in a certain period of time, where the buyer and seller are in specific countries, and the product has parts manufactured in yet another country. Yet, long running times for such analytical queries, even in leading, commercial-level database systems, are still among the major challenges to overcome.

Now Prof Ke Yi, an expert in database systems and algorithms, has solved a problem that has taxed the community for over 15 years, enabling responses to queries to be given in seconds rather than minutes or hours. Prof Yi and his team's novel algorithm allows the database to return approximate results in a very short time, and continue to improve their accuracy as more time is spent. Working on

datasets of up to 100 gigabytes (and potentially larger), their “Wander Join” algorithm has achieved better sampling through random walks, returning results with the same accuracy (for example, 95% confidence and 1% error) in one-hundredth of the time compared with prior solutions using the same hardware.



“
In the early days,
proving a conjecture and
having a theorem with
my name on it was my dream.
Now seeing my algorithm
used in practice is
more satisfying”

PROF KE YI
Associate Professor of Computer Science
and Engineering

Random walk is a technique that has been used to solve problems in many fields. Google's successful search engine is based on this idea, enabling it to prioritize the most authoritative pages related to the keywords being searched with a high degree of accuracy out of millions of web pages. In Prof Yi's research, this theoretical idea is successfully applied for the first time to the completely different scenario involved in approximate query processing. In addition, the algorithm has been integrated into an open source database to demonstrate its viability in the real world and not just in academia, bringing the sought-after goal of interactive data analysis a step closer.

Assisting such original insight is Prof Yi's unusual status as a member of both theoretical computer science and database faculty groups at HKUST, allowing him to bring knowledge from the two areas together to address practical problems. International recognition for the work of Prof Yi and his team includes winning the 2015 ACM SIGMOD Best Demonstration Award and receiving the 2016 ACM SIGMOD Best Paper Award. The work was carried out in collaboration with the University of Utah.

Making Data Centers Communicate Faster

Enter a single web search query and it can set in motion communication between thousands of physical machines in a data center as the machines quickly retrieve and collect the information corresponding to your keywords. A major goal of Prof Kai Chen and his research group is to accelerate such communication (data flow) between machines to help deliver the cloud and big data applications that now facilitate our way of life, including

through innovative solutions that could be directly implemented.

KARUNA optimizes cloud application performance by delivering the first mix-flow (data flows with and without deadlines) scheduling solution for data centers. Unlike existing solutions, KARUNA maximizes the deadline meet rate of data flows with deadlines while minimizing the transmission time of data flows without deadlines. This is achieved by prioritizing deadline flows while controlling their sending rates,



“Data centers provide the infrastructure for big data and cloud computing. Our goal is to make data center communication faster”

PROF KAI CHEN

Associate Professor of Computer Science and Engineering

performance to be equivalent to solutions requiring applications to be adapted.

Prof Chen has also worked closely with major companies. Collaborations with Huawei have led to technologies for Software Defined Networks (SDN), prototypes that use machine learning for efficient communication, and patents. Prof Chen received Huawei's inaugural Distinguished Collaborator Award in 2016. At Tencent, the HKUST team has contributed to new-generation machine learning system Angel by designing an efficient data flow scheduling scheme that can improve the machine learning algorithm convergence time by up to 90%. The overall performance of Angel is 70 times faster than previous systems tested. Tencent has deployed it to support advertisement and video recommendation services.

Making these advances possible is the 100-plus machine data center that Prof Chen and his team have built from scratch, providing an essential in-house testbed at HKUST to try out the feasibility of a proposed solution. Once outcomes reach the required performance levels, large-scale simulations can be used to demonstrate scalability. Meanwhile, Prof Chen is setting his sights on further frontier work, including optical networking and artificial intelligence(AI)-enabled networking, currently undergoing testing at HKUST.

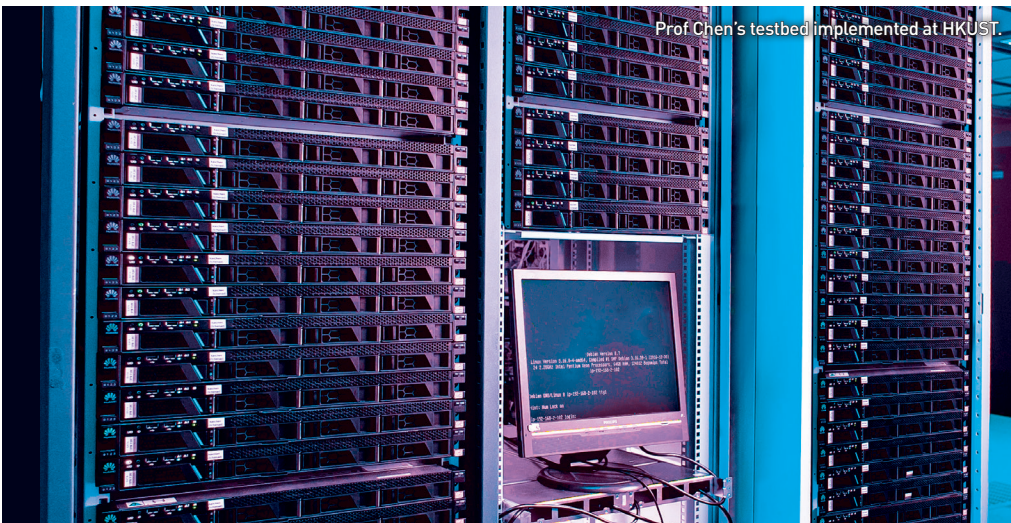
search engines, financial services, social networking, and many others.

Like Prof Ke Yi's work on databases, Prof Chen's approach to networked system design is theoretically significant and also practical. He seeks to achieve very high throughput and low latency (faster speed at the millisecond/microsecond level) while not requiring industry to make hard-to-adopt changes to applications or customize hardware. He is proving successful. In 2016, his team became the first in Hong Kong to have first-author papers accepted for the eminent ACM SIGCOMM conference. In these two papers, Prof Chen's group designed two systems, KARUNA and CODA, that answered key problems

using the remaining bandwidth to schedule non-deadline flows. The system does not require prior knowledge of data size or impractical switch hardware modifications, filling an important gap in data center flow scheduling.

CODA is a system that can automatically identify and schedule coflows (a collection of parallel flows sharing a common performance goal) without requiring changes to applications – an impractical requirement of other coflow-based solutions. This is achieved by employing a machine learning algorithm to rapidly identify coflows, complemented by a coflow scheduler which is tolerant of identification errors. Testbed and large-scale simulations showed CODA's

Prof Chen's testbed implemented at HKUST.



WHO'S WHO IN THE ARTS

Cutting-edge mathematical forensics at HKUST is bringing quantitative insights to art and literature



Two pictures, both thought to be the work of Raphael. How can mathematics help to ascertain if they really are by the famous painter? This intriguing area is among the domains of Prof Yang Wang, a specialist in mathematical forensics and stylometry, namely quantitative analysis of artistic or literary style.

To ascertain whether a specific painter or author created a work, he and his team are using the combined reach of machine learning and randomization theory to add to traditional methods of authentication, such as observational expertise and connoisseurship. For example, in visual art analysis, brush stroke measurements, texture models, fractal models, and color palette can be compared with other works by the same painter. In literary stylometry, average length of sentences, synonym pairs, and frequency of words, among other features, can be analyzed.

“The computing power for what we do was available 15 years ago, but people simply didn’t think about doing such analysis,” Prof Wang said. “What has made the difference is the increasing sophistication of techniques in statistical analysis, pattern recognition, signal processing, and machine learning that have opened the way for almost everything to be looked at from the perspective of big data and data analytics.”

Prof Wang first became interested in stylometry in 2009 when the Metropolitan Museum of Art in the US asked him and a collaborator to examine a corpus of drawings by the great Flemish artist Pieter Bruegel the Elder (1525-1569) and famous imitations to see whether mathematical techniques could provide insight into identifying forgeries. Prof Wang’s team later published their results in the international journal, *IEEE Transactions of Pattern Recognition*

“

I’m not an artist or historian, but this interplay of mathematics being applied to areas where few thought it could be applied is really exciting

”

PROF YANG WANG
Chair Professor of Mathematics,
Dean of Science

and Machine Intelligence. Since joining HKUST in 2014, he has helped his team members to develop more novel mathematical techniques in stylometry. These include a new randomization technique to analyze challenging open cases in authorship authentication, where the authorship in question is not limited to a small group of “suspects”.

His team’s research has been published in or submitted to peer-reviewed journals, such as *Applied and Computational Harmonic Analysis*, *Adaptive Data Analysis*, and *Signal Processing*.

In addition, Prof Wang sees great possibilities for mathematics to extend from arts and culture into other areas in social science and humanities in the future. “We have been looking at authorship stylometry but such work could easily move into other areas, such as computational rhetoric, sentiment analysis, text mining, image classification, even government surveillance and fingerprint analysis. We view this as an interdisciplinary area with far-reaching impact down the road.”

Is This a Raphael?



A drawing purported to be a Raphael.

The drawing above is part of a private collection. Was it drawn by Raphael? To conduct stylometric analysis, HKUST researchers first extracted quantitative features from the drawing using wavelet decomposition and scattering transform. Similar features were extracted from a number of drawings known to be genuine Raphael works or forgeries provided by the collector. Features were then compared. Although such analysis does not draw definitive conclusions, their results showed that the drawing’s style was consistent with a Raphael. The collector also provided some drawings of unknown provenance that bore a strong resemblance to those by Raphael. HKUST researchers showed stylistically they were inconsistent with Raphael.



Two works already established to be by the famous painter.